Feature Expansion for Sentiment Analysis in Twitter

Erwin B. Setiawan, Dwi H. Widyantoro, Kridanto Surendro
School of Electrical Engineering and Informatics, Institut Teknologi Bandung
Jl Ganesha No 10, Bandung, Indonesia
Email: erwinbudisetiawan@telkomuniversity.ac.id, dwi@stei.itb.ac.id, endro@stei.itb.ac.id

Abstract—The community's need for social media is increasing, since the media can be used to express their opinion, especially the Twitter. Sentiment analysis can be used to understand public opinion a topic where the accuracy can be measured and improved by several methods. In this paper, we introduce a hybrid method that combines: (a) basic features and feature expansion based on Term Frequency–Inverse Document Frequency (TF-IDF) and (b) basic features and feature expansion based on tweet-based features. We train three most common classifiers for this field, i.e., Support Vector Machine (SVM), Logistic Regression (Logit), and Naïve Bayes (NB). From those two feature expansions, we do notice a significant increase in feature expansion with tweet-based features rather than based on TF-IDF, where the highest accuracy of 98.81% is achieved in Logistic Regression Classifier.

Keywords—sentiment analysis; feature expansion; twitter

I. INTRODUCTION

At the end of January 2018, We Are Social and Hootsuite, released data on the number of Internet users and social media in the world [10]. Based on that data, internet users in the world has reached 4 billion, previously, 3.8 billion. Until the first quarter-2017, Twitter users worldwide reached 328 million, an increase of about 14% over the same period in the previous year. From the data released by Twitter Indonesia at the end of 2016, it was noted that 77 percent of Twitter users in Indonesia are active users. In addition, Twitter users in Indonesia are also among the fussiest. This can be seen from the number of tweets generated throughout 2016 which reached 4.1 billion tweets. The number of Twitter users in Indonesia is a promising market, including in five major worlds [11]. It is not surprising that in various fields, e.g., in economics, producers are competing to manage this great potential for their products on the market.

Sentiment analysis is part of opinion mining [1]. Sentiment analysis is the process of understanding, extracting and processing textual data automatically to get sentiment information contained in an opinion sentence. The magnitude of the influence and benefits of sentiment analysis led to research or application of sentiment analysis growing rapidly, even in America approximately 20-30 companies that focus on sentiment analysis services [1].

Less applications and methods of sentiment analysis developed for twitter speak Indonesian. This sentiment analysis research was conducted to find out the sentiments of a tweet on Twitter by using the approach in machine learning that is Naïve Bayes (NB), Logistic Regression (Logit) and Support Vector Machine (SVM) devoted to Twitter in Indonesian language with various features. Variations that basic features used in this research are unigram, bigram, trigram, POS (Part-of-Speech) Tags, POS Bigram, POS Trigram, and Line Length. In this research, we introduce basic features with feature expansion by using TF-IDF and feature expansion by using tweet-based features. The aim of the paper is to find out the best sentiment analysis models that happened based on accuracy value.

The rest of the paper is organized as follows. Section II discusses issues related to sentiment analysis techniques. Section III describes system model of sentiment analysis on Twitter. Section IV provides the experimental and analysis, followed by the conclusion in section V.

II. RELATED WORK

The related research on sentiment analysis is abundant. M. S. Neethu and R. Rajasree [2] measured the accuracy of classification process using various classifiers such as Nave Bayes, Maximum Entropy, Support Vector Machine, and Ensemble classifiers. They obtained an accuracy of 90% whereas Naïve Bayes has 89.5%. The feature space used included feature expansion by using tweet-based features, i.e., hashtags and emoticons marks in conjunction with features like unigram.

Celikyilmaz *et al.* [3] used Naïve Bayes, SMO, SVM and Random Forest to classify Twitter data, with F-scores that are relatively 10% better than a classification baseline that uses raw word n-gram features. The features considered by classifiers were pronunciations of words, polarity lexicon from tweets, and extract a set of features based on this lexicon.

Barbosa *et al.* [4] proposed a 2-step sentiment analysis method for classifying tweets. The first step, they classified tweets into subjective and objective and the second phase, the subjective tweets into positive or negative. They use SVM classifier. The feature space used included tweet-based features (retweets, hashtags, link, punctuation) and exclamation marks in conjunction with features like prior polarity of words and POS.

Bahrainian and Dengel [5] used SVM, Maximum Entropy, and Naïve Bayes to examine the performance based on the unigram feature set and compare with a hybrid method that is a combination of the usage of sentiment lexicons with a

978-1-5386-8402-3/18/$31.00 ©2018 IEEE

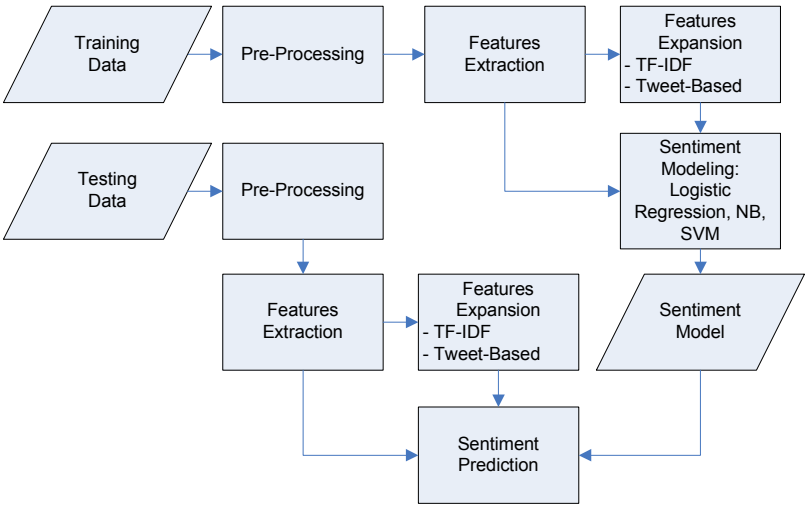

Fig. 1. Sentiment Analysis Model

machine learning classifier for polarity detection of subjective texts in the consumer-products domain. The experimental results indicate that our hybrid method outperforms all other mentioned methods.

In this paper, we introduce a different approach compared to previous research are scope twitter's account of the Indonesian-language, a hybrid method that is a combination between basic features with feature expansion based on TF-IDF and feature expansion based on tweet-based features. We use several classification algorithms, such as Naïve Bayes (NB), SVM, and Logistic Regression (Logit).

III. Sentiment Analysis Model and The ProposedTechnique

Sentiment Model is shown in Fig. 1. Dataset is divided into two, i.e., training data and testing data. Each data is done pre-processing, feature extraction and feature expansion. Furthermore, the feature extraction or feature expansion results for the training data as input to the modeling process sentiment, the results of this modeling are used to predict the testing data. Last is expected to get Sentiment Class with excellent accuracy.

A. Crawling Data

Crawling data on twitter is a process to retrieve or download data from twitter server with the help of Application Programming Integration (API) twitter either in the form of user data or tweet data. The twitter data do be the reference data of this research. Twitter data is divided into two data, training data and testing data.

B. Pre-Processing

In the step, we use is pre-processing such as case folding, tokenization, stopwords removal, and stemming. We do pre-processing twitter data automatically with the application that we developed in previous research [6].

C. Feature Extraction

Feature extraction is process of taking the feature of a tweet that can describe the characteristics of the tweet. In this phase, we use basic features such as unigram, bigram, trigram, POS (Part-of-Speech) Tags, POS Bigram, POS Trigram, and Line Length.

D. Feature Expansion

In this research, we use two feature expansions. Two feature expansions are expansion using TF-IDF and expansion with tweet based feature. Two feature expansions use to improve of accuracy from analysis sentiment.

For TF-IDF as commonly used TF-IDF formula in [7]. The weight of word k in a tweet T is calculated as follows:

$$W_{ki} = tf_{kT} * \log\left({}^{N_T}/_{n_k}\right) \quad (1)$$

Where tf_{kT} is the appearance frequency of word k in the tweet T, N is the number of all the collection of tweets that has been used and n_k is s the number of tweets containing the word.

For feature expansion by using TF-IDF, after we get a unique word that weights for each sentiment (negative, positive, and neutral), the features are as in the research that has been done in [8]. The representation of a 1_tfidf tweet vector of results can be illustrated as follows, suppose the feature vector encodes the appearance of word in the following order: “good” is 1-top feature in positive sentiment, “bad” is 1-top feature in negative sentiment, and “new” is 1-top feature in neutral sentiment, respectively. A tweet containing “He is a good man” will be represented as {1,0, 0}.

For feature expansion with tweet-based features, we get 25 features. These features are taken when crawling data and shown in Table I.

E. Classification Algorithm

In the research, we use three classifiers used to create classification model, i.e., Naïve Bayes (NB), Logistic Regression (Logit), and Support Vector Machine (SVM).

Naïve Bayes (NB) is one of the classification methods based on Bayes theorem by using probability and statistical techniques. The Naïve Bayes algorithm predicts future opportunities based on prior experience with the main characteristic is a very strong assumption of independence from each condition [9].

Support Vector Machine (SVM) is a classification method that includes supervised learning that analyzes data and recognizes the data patterns. SVM classification is the process of finding the best hyperplane line that separates two

classes in the input space where the maximum distance or margin to the nearest pattern class points or support vector [9].

Logistic Regression (Logit) is a mathematical model that can be used to describe the relationship between variable X and the dependent variable (Y). In Logistic Regression variable x which is predicted is a function of the probability that an object will be in one category (Y) [9].

As shown in Fig. 1, the algorithms of classification are used to create sentiment model during training phase. The sentiment analysis model is then used to classify of new tweets, using the same algorithms as used to create the classification model.

IV. EXPERIMENTS AND ANALYSIS

There are 3 (three) objectives of experiment, that is, first to know comparison each labeling technique, second to know influence of feature expansion of TF-IDF and of tweet-based features, and third to know influence of tweet-based features. The classification accuracy is defined as the percentage of correctly classified instances using 10-fold cross validation.

TABLE I. 25 FEATURES OF TWEET-BASED

No	Feature	Description
1	#char	The number of characters from a tweet
2	#emot_happy	The number of emoticons containing expressions of happy
3	#emot_sad	The number of emoticons containing expressions of sad
4	#hashtag	Number of hashtags from a tweet
5	#mention	Number of mentions from a tweet
6	#url	Number of urls from a tweet
7	check_spam	To find out whether in a tweet there are words that are included in the spam list.
8	has_happy	The presence of emoticons that contain expressions of happy
9	has_sad	The presence of emoticons that contain expressions of sad
10	is_favorited	Represented by a small star icon next to a Tweet, are most commonly used when users like a Tweet
11	is_hashtag	The presence of hashtag (#) in a tweet
12	is_mention	The presence of mention(@) in a tweet
13	is_retweet	The presence of retweet (RT) in a tweet
14	is_url	The presence of URL in a tweet
15	tot_negative	The number of negative words from a tweet
16	tot_positive	The number of positive words from a tweet
17	tot_sentiment	The number of sentiment words from a tweet
18	tot_word	The number of words from a tweet
19	lenght_tweet	Character length or word length of a tweet
20	ratioNegNumtweet	The ratio of the number of negative sentiments to the number of words in a tweet
21	ratioPosNumtweet	The ratio of the number of positive sentiments to the number of words in a tweet
22	ratio_char_tot_word	The ratio of the number of character to the number of words in a tweet
23	ratio_char_lenght_tweet	The ratio of the number of positive sentiments to the length of words in a tweet
24	retweet_counted	The number of users who retweet a tweet
25	source	The device that used to share a tweet. Grouped into two, via smartphone or PC Client.

A. Data Set and Labeling

We use the same dataset used in [8], which contains 19.401 tweets in Bahasa Indonesia. There are tree data labeling, such as

- Manual labeling
  This labeling involves 20 students. Labeling results can be seen in Table II.

TABLE II. MANUAL LABELING DISTRIBUTION

Label	Sum	Percentage
Positive	8078	41.64%
Negative	2611	13.46%
Neutral	8712	44.90%
Total	19401	

- Labeling by system
  In this labeling, we create a corpus that contains a list of words of sentiment, 354 words, then conducted a survey of each word to get negative, positive and neutral sentiment. This sentiment is obtained by searching for words that belong to negative, positive and neutral groups. Some sample data can be seen in Table III. Labeling results can be seen in Table IV.

TABLE III. EXAMPLE 10 SENTIMENT SURVEY

No	Word	Positive (%)	Negative (%)	Neutral (%)	Weight
1	buruk	0	78.3	21.7	3
2	jelek	0	78.3	21.7	3
3	lama	4.3	30.4	65.3	0
4	lamban	4.3	78.3	17.4	3
5	lambat	13	52.2	34.8	1
6	baik	82.6	0	17.4	4
7	berani	82.6	0	17.4	4
8	benar	82.6	0	17.4	4
9	sudah	56.5	0	43.5	1
10	Ayo	65.2	4.3	30.5	2

TABLE IV. LABELING BY SYSTEM DISTRIBUTION

Label	Sum	Percentage
Positive	9798	50.50%
Negative	1645	8.48%
Neutral	7958	41.02%
Total	19401	

- Labeling by system plus emoticon
  After labeling by system, the next step we do with emoticons. There are two cases handled in this label, first, if the neutral system but positive emoticons then the result is positive, and secondly, if the neutral system but negative emoticons then the result is negative. Labeling results can be seen in Table V.

TABLE V. LABELING BY SYSTEM PLUS EMOTICON DISTRIBUTION

Label	Sum	Percentage
Positive	9797	50.50%
Negative	1655	8.53%
Neutral	7949	40.97%
Total	19401	

B. Comparison Labeling

As shown in Table VI, labeling by system provides improved accuracy in all conditions. The highest accuracy was achieved at 83.94% on the basic features of Unigram+Bigram+Trigram+Post on logistic regression classifier.

TABLE VI. MANUAL VS SYSTEM LABELING (CORPUS SENTIMENT)

Condition			%		
			NB	Logit	SVM
Unigram	39100 features	System	67.75	83.42	81.29
		Manual	56.68	58.23	56.12
Bigram	37096 features	System	67.33	82.71	80.98
		Manual	56.47	58.13	56.16
Uni+Bi+Trigram+POS	50779 features	System	**69.01**	**83.94**	**82.05**
		Manual	57.65	58.65	56.60
All	52459 features	System	68.77	83.68	82.05
		Manual	57.50	58.65	56.46

Likewise labeling by system when compared with labeling by system with emoticons. The highest results are still achieved labeling by system. Except one cell of SVM classifier, here is increase accuracy that is equal to 0.14% The results can be seen in Table VII.

TABLE VII. LABELING BY SYSTEM VS BY SYSTEM WITH EMOTICON

	50779 features Uni+Bi+Trigram+POS (%)	
	System	System with emoticon
NB	69.01	68.92 (-0.13)
Logit	83.94	83.91 (-0.03)
SVM	82.05	**82.17 (+0.14)**

In this section, experiments are also performed for labeling by system with some basic features and their combinations, their performances shown in Table VIII. The basic features used include Unigram, Bigram, Trigram, POS Tags, and Line Length. The highest accuracy was seen in the combination of basic features consisting of Unigram, Bigram, Trigram and POS Tags, with accuracy of 69.01% for NB, 83.94% for Logit and 82.05 % for SVM.

TABLE VIII. LABELING BY SYSTEM WITH BASIC FEATURE

Basic Feature	#Features	(%)		
		NB	Logit	SVM
Trigram	33544	66.85	82.28	80.13
Bigram	37096	67.33	82.71	80.98
Unigram	39100	67.75	83.42	81.29
POS Tags (POS)	39448	67.60	83.23	81.21
Unigram+Line Length	39101	67.66	83.40	81.28
Bigram+POS	43741	68.07	83.49	81.66
Unigram+POS	45745	68.39	83.69	81.68
Uni+Bigram+POS+Line Length	50039	68.77	83.92	82.01
Uni+Bi+Trigram+POS	50779	**69.01**	**83.94**	**82.05**
Uni+Bi+Trigram+POS+ Line Length (All)	52459	68.77	83.68	**82.05**

C. Labeling by System with Feature Expansion

The baseline row describes the results without performing feature expansion that is accuracy labeling by system using basic features Unigram, Bigram, Trigram and POS Tags. In Table IX, we can see that the effect of feature expansion has significant effect on SVM classifier, the accuracy has increased. From two feature expansions, we do notice a significant increase in feature expansion with tweet-based features, the highest accuracy of 98.81 % is achieved in logistic regression classifier.

TABLE IX. LABELING BY SYSTEM VS BY SYSTEM WITH FEATURE EXPANSION

Condition	(%)		
	NB	Logit	SVM
Baseline (B)	69.01	83.94	82.05
B +1_tfidf	68.9 (-0.09)	83.9 (-0.06)	82.2(**+0.14**)
B+5_tfidf	68.9 (-0.08)	83.9 (-0.06)	82.2(**+0.16**)
B+10_tfidf	68.9 (-0.08)	83.9 (-0.06)	82.2(**+0.19**)
B +20_tfidf	68.9 (-0.08)	83.9 (-0.04)	82.3 (**+0.25**)
B +50_tfidf	70.5 (**+2.09**)	83.9 (-0.02)	82.1 (**+0.04**)
B +100_tfidf	68.9 (-0.08)	83.9 (**+0.02**)	82.3 (**+0.34**)
B+Tweet-based	82.4 (**+19.5**)	**98.81 (+17.7)**	92.1(**+12.2**)

D. Influence of tweet-based features

As shown in Table X, the features that positively affect the accuracy are baseline plus tot_sentiment, tot_positive, #emot_happy, #emot_sad, ratioNegNumtweet, and RatioPosNumtweet. While the negative effect is the feature baseline plus #hashtag, #url, tweet_length, retweet_counted and source. The other features (12 features) provide mixed results for three classifiers.

TABLE X. INFLUENCE OF TWEET–BASED FEATURES

Condition	%		
	NB	Logit	SVM
Baseline (B)	69.01	83.94	82.05
B+all	**82.4 (+19.5)**	**98.8 (+17.7)**	**92.1 (+12.2)**
B+#emot_happy	**69.1 (+0.13)**	**84.9 (+1.19)**	**83.5 (+1.78)**
B+#emot_sad	**69.0 (+0.00)**	**83.9 (+0.06)**	**82.0 (+0.03)**
B+#hashtag	68.9 (-0.03)	83.8 (-0.06)	81.9 (-0.14)
B+#url	68.9 (-0.04)	83.9 (-0.02)	82.0 (-0.03)
B+tot_negative	**69.9 (+1.36)**	**85.9 (+2.41)**	**82.9 (+1.14)**
B+tot_positive	**74.2 (+7.63)**	**93.0 (+10.8)**	**89.4 (+9.05)**
B+tot_sentiment	**79.2 (+14.7)**	**98.7 (+17.7)**	**88.8 (+8.27)**
B+ length_tweet	68.8 (-0.22)	83.9 (-0.02)	82.0 (-0.03)
B+ratioNegNumtweet	**70.4 (+2.10)**	**84.5 (+0.76)**	**88.8 (+8.27)**
B+ratioPosNumtweet	**76.5 (+10.9)**	**91.0 (+8.41)**	**85.6 (+4.32)**
B+retweet_counted	68.8 (-0.19)	82.0 (-2.31)	82.0 (-0.04)
B+source	69.0 (-0.01)	83.9 (-0.05)	82.0 (-0.01)

V. CONCLUSION

We have introduced a hybrid method combining basic feature and feature expansion for improving the accuracy of sentiment analysis in Twitter. We have trained SVM, Logit, and NB to observe the accuracy of sentiment analysis using a series of computation. Expansion features can be used and is proven to increase the accuracy of sentiment analysis. From two feature expansions, we do notice a significant increase in feature expansion with tweet-based features, where the highest accuracy of 98.81% is achieved using logistic regression classifier.

Among tweet-based features, we found that the features that affecting positively the accuracy are tot_sentiment, tot_positive, #emot_happy, #emot_sad, ratioNegNumtweet, and RatioPosNumtweet. While the negative effect is coming from the feature #hashtag, #url, length_tweet, retweet_counted and source.

ACKNOWLEDGMENT

The author would like to thank to PDD Hibah Dikti 2018, STEI ITB and BPPDN RISTEKDIKTI for the support to this research.

REFERENCES

[1] B. Liu, "Sentiment Analysis and Subjectivity," Handb. Nat. Lang. Process., no. 1, hal. 1–38, 2010.

[2] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2013.

[3] A. Celikyilmaz, D. Hakkani-Tür, J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," 2010 IEEE Work. Spok. Lang. Technol. SLT 2010 - Proc., hal. 79–84, 2010.

[4] L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," Coling, no. August, hal. 36–44, 2010.

[5] S. A. Bahrainian and A. Dengel, "Sentiment Analysis using sentiment features," Proc. - 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IATW 2013, vol. 3, hal. 26–29, 2013.

[6] E. B. Setiawan, D. H. Widyantoro, K. Surendro, "Detecting Indonesian Spammer on Twitter," 6th Int. Conf. Inf. Commun. Technol. (ICOICT), May 3-4 , 2018, no. 10, 2018.

[7] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," J. Doc., vol. 60, no. 5, hal. 503–520, 2004.

[8] E. B. Setiawan, D. H. Widyantoro, K. Surendro, "Feature Expansion using Word Embedding for Tweet Topic Classification," Inf. Commun. Technol. (ICoIC7), 2017 5th Int. Conf. Malacca City, Malaysia, 2016.

[9] T. M. Mitchell, Machine Learning. McGraw-Hill Science, 1997.

[10] https://www.merdeka.com/teknologi/media-sosial-merajai-penggunainternet-di-dunia.html, accessed June 2018

[11] http://www.beritasatu.com/digital-life/428591-indonesia-masuk-limabesar-pengguna-twitter.html, accessed June 2018